On Testing for Informative Selection in Survey Sampling

(plus Some Estimation)

Jay Breidt
Colorado State University

Survey Methods and their Use in Related Fields
Neuchâtel, Switzerland
August 23, 2018

Joint work with various people, acknowledged as we go.

Goal: Inference for the distribution of Y

- Finite population $U = \{1, 2, \ldots, N\}$
- Random variables $\{Y_k : k \in U\}$ are independent and identically distributed
- Observe the realized values not for all of U, but only a random subset:

$$\{y_k : k \in s \subset U\}$$

- Goal is inference on the distribution of Y, or some of its characteristics
- Concerned about effect of selection of $s \subset U$ on inference

Sample membership indicators

- Define sample membership indicators I_k, where

$$I_k = \begin{cases} 1 & \text{if } k \in s \\ 0 & \text{otherwise} \end{cases}$$

- If the selection **is** designed/controlled, the event $\{k \in s\}$ may depend on Y_k
- If the selection **is not** designed/controlled, the event $\{k \in s\}$ may depend on Y_k
- Probability of selection, in general, may depend on Y_k

Inclusion probabilities

- To allow probability of selection to depend on Y_k, make it random
- Inclusion probability is the realization of random variable Π_k that may depend on Y_k:

$$\begin{aligned}\pi_k &= \mathbb{P}\left[I_k = 1 \mid Y_k = y_k, \Pi_k = \pi_k\right] \\ &= \mathbb{E}\left[I_k \mid Y_k = y_k, \Pi_k = \pi_k\right]\end{aligned}$$

Examples with explicit dependence on Y_k

- Cut-off sampling: $\pi_k = \rho(y_k)\mathbb{1}_{\{y_k > \tau\}}$.
- Case-control study (binary Y):

$$\pi_k = \begin{cases} 1, & \text{for disease cases } (y_k = 1) \\ \rho < 1, & \text{for non-disease controls } (y_k = 0) \end{cases}$$

- Choice-based sampling (categorical Y):

$$\pi_k = \sum_{j=1}^{J} \rho_j \mathbb{1}_{\{y_k = j\}}.$$

- Adaptive sampling, quota sampling, endogenous stratification, ...

Length-biased sampling

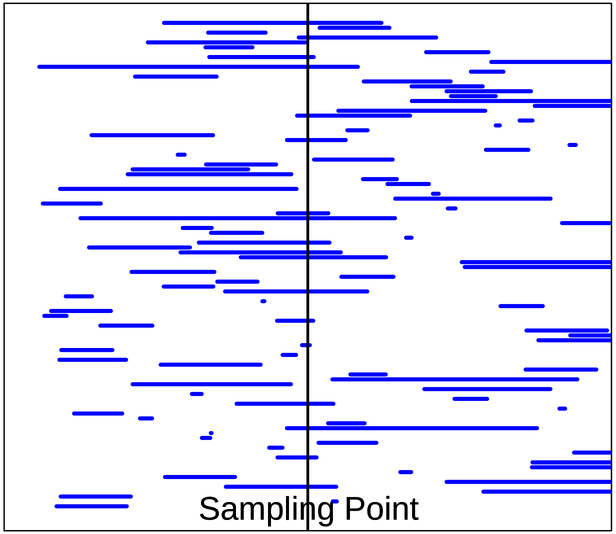
- Length-biased sampling: $\pi_k \propto y_k > 0$
- Good design for y_k tries to be length-biased
- Why? For fixed size design,

$$
\begin{aligned}
\text{Var}\left(\sum_{k \in s} \frac{y_k}{\pi_k} \middle| \boldsymbol{Y}_U = \boldsymbol{y}_u, \boldsymbol{\Pi}_U = \boldsymbol{\pi}_U \right) &= -\frac{1}{2} \sum_{j,k \in U} \Delta_{jk} \left(\frac{y_j}{\pi_j} - \frac{y_k}{\pi_k} \right)^2 \\
&= -\frac{1}{2} \sum_{j,k \in U} \Delta_{jk} \left(\frac{y_j}{c y_j} - \frac{y_k}{c y_k} \right)^2 \\
&= 0
\end{aligned}
$$

- Unbiased estimator with zero variance!

Length-biased sampling: $\pi_k \propto y_k$

y = textile fiber length (Cox, 1969), intercepted individual's time spent at recreational site, size of sighted wild animal, lifetime of marked-recaptured individual, disease latency period,. . .



Implicit dependence on Y_k

- Often, Π_k does not depend explicitly on Y_k, but Y_k has predictive power for Π_k
- Consider parametric empirical models:

$$\mathbb{E}\left[\Pi_k \mid Y_k = y_k\right] = \mu(y_k; \xi),$$

where ξ are nuisance parameters with respect to Y

- Or consider nonparametric empirical models:

$$\mathbb{E}\left[\Pi_k \mid Y_k = y_k\right] = \mu(y_k),$$

The effect of selection

- Parametric model for average inclusion probability:

$$\mathbb{E}\left[\Pi_k \mid Y_k = y_k\right] = \mu(y_k; \xi)$$

- Relevant distribution of observed Y_k is

$$f(y \mid I_k = 1) = \frac{\mu(y;\xi)}{\int \mu(y;\xi) f(y)\, dy} f(y) =: \rho(y;\xi) f(y),$$

in which the denominator depends on f

- If μ does not depend on y, then

$$f(y \mid I_k = 1) = \frac{\mu(\xi)}{\mu(\xi) \int f(y)\, dy} f(y) = f(y)$$

- Suppose Y_k iid $\mathcal{N}(\theta, \sigma^2)$
- Further suppose:

$$\Pi_k \mid (Y_k = y_k) \sim \log \mathcal{N}\left(\xi_0 + \xi y_k, \tau^2\right)$$

$$\mathbb{E}\left[\Pi_k \mid Y_k = y_k\right] = \exp\left(\xi_0 + \xi y_k + \frac{\tau^2}{2}\right)$$

- Then it is easy to show that

$$Y_k \mid (I_k = 1) \sim \mathcal{N}\left(\theta + \xi\sigma^2, \sigma^2\right),$$

so sample mean will be biased and inconsistent for θ

Application to a textbook survey

- Simulated data from Fuller (2009, Ex. 6.3.1) following Korn and Graubard (1999, Ex. 4.3-1) for 1988 National Maternal and Infant Health Survey
- Conducted by US National Center for Health Statistics
- Goal: study factors related to poor pregnancy outcome
- Design: nationally-representative stratified sample from birth records, with oversampling of low-birthweight infants
  - complex survey: stratified, unequal-probability

Selection for NMIHS

- Let $U =$ all US live births in 1988
- Let $Y_k =$ gestational age, strongly related to birthweight
- Suppose Y_k iid $\mathcal{N}(\theta, \sigma^2)$
- Inclusion probability in NMIHS depends on birthweight, hence Y_k is predictive:

$$\mathbb{E}\left[\Pi_k \mid Y_k = y_k\right] = \exp\left(\xi_0 - 0.175 y_k + \frac{\tau^2}{2}\right)$$

- Greater gestational age $\Rightarrow$ less likely to be sampled

- By previous computation, negative bias in the unweighted sample mean:

$$Y_k \mid (I_k = 1) \sim \mathcal{N}\left(\theta - 0.175\sigma^2, \sigma^2\right),$$

```
> svymean(~GestAge, birth.design)
           mean     SE
GestAge 39.138 0.0941
> # Unweighted minus weighted:
> mean(birth$GestAge) - svymean(~GestAge, birth.design)
        -2.2114
```

- Here we used classical design-based techniques to deal with effects of selection

- Provided $\pi_k > 0$ for all $k \in U$ plus additional mild conditions,
$$\widehat{\theta}_{\mathrm{HT}} = \frac{1}{N} \sum_{k \in U} y_k \frac{I_k}{\pi_k}$$
is unbiased and consistent for finite-population average:
$$\mathbb{E}\left[\frac{1}{N} \sum_{k \in U} y_k \frac{I_k}{\pi_k} \,\middle|\, \boldsymbol{\pi}_U, \boldsymbol{y}_U \right] = \frac{1}{N} \sum_{k \in U} y_k = \theta_N$$
- Consistency for θ then follows by chaining argument:
$$\widehat{\theta}_{\mathrm{HT}} - \theta = \left(\widehat{\theta}_{\mathrm{HT}} - \theta_N \right) + (\theta_N - \theta) = \text{small} + \text{smaller}$$

Horvitz-Thompson plug-in principle: explicit

- If finite population parameter can be written explicitly as

$$\theta_N = \vartheta\left(\sum_{k\in U} y_k^{(1)}, \ldots, \sum_{k\in U} y_k^{(p)}\right)$$

for some smooth map $\vartheta(\cdot)$, then

$$\widehat{\theta}_{\mathrm{HT}} = \vartheta\left(\sum_{k\in U} y_k^{(1)}\frac{I_k}{\pi_k}, \ldots, \sum_{k\in U} y_k^{(p)}\frac{I_k}{\pi_k}\right)$$

is consistent and asymptotically design-unbiased for θ_N

Horvitz-Thompson plug-in principle: implicit

- If a finite population parameter can be written as solution to a population-level estimating equation,

$$\theta_N \text{ solves } \mathbf{0} = \boldsymbol{\varphi}\left(\sum_{k\in U} y_k^{(1)}, \ldots, \sum_{k\in U} y_k^{(p)}; \theta\right),$$

then HT plug-in estimator is obtained by solving weighted sample-level estimating equation:

$$\widehat{\theta}_{\mathrm{HT}} \text{ solves } \mathbf{0} = \boldsymbol{\varphi}\left(\sum_{k\in U} y_k^{(1)} \frac{I_k}{\pi_k}, \ldots, \sum_{k\in U} y_k^{(p)} \frac{I_k}{\pi_k}; \theta\right)$$

Pseudo-likelihood estimation

- If estimating equation uses the population-level score,

$$\mathbf{0} = \frac{\partial}{\partial \theta} \sum_{k \in U} \ln f(y_k; \theta) \bigg|_{\theta = \theta_N},$$

  then θ_N are population-level MLE's

- If it uses the weighted sample-level score,

$$\mathbf{0} = \frac{\partial}{\partial \theta} \sum_{k \in U} \ln f(y_k; \theta) \frac{I_k}{\pi_k} \bigg|_{\theta = \widehat{\theta}_{\mathrm{HT}}},$$

  then $\widehat{\theta}_{\mathrm{HT}}$ are maximum pseudo-likelihood estimators

HT plug-in principle plus chaining argument

- Combining plug-in and chaining argument:
  - Link 1: for the superpopulation model parameter θ, define a corresponding finite population parameter θ_N
  - Link 2: estimate θ_N by $\widehat{\theta}_{\mathrm{HT}}$ using HT plug-in principle
- Typically,

  $$\widehat{\theta}_{\mathrm{HT}} - \theta = \left(\widehat{\theta}_{\mathrm{HT}} - \theta_N\right) + (\theta_N - \theta) = O_p\left(n^{-\alpha}\right) + O_p\left(N^{-\alpha}\right)$$

  where $n << N$, so ignore the second component
- Use design-based methods to estimate the variance of the first component, ignoring the second

Options for dealing with selection

- **Default Option:** Assume informative selection
  - use HT plug-in and chaining
  - simple and readily available in software
  - design-based option is not usually the most efficient
- **Other Options:** Test for informative selection
  - if no evidence of selection effects, proceed with fully-efficient likelihood-based methods
  - if evidence of selection effects, proceed with likelihood-based procedures that account for effects of selection

- **Pseudo-likelihood:** easy but least efficient
- **Full likelihood:** most efficient, often impractical
  - in general, joint distribution of all observed Y_k, I_k, Π_k
  - with no selection, joint distribution of Y_k only
- **Sample likelihood:** treat $\{Y_k\}_{k \in s}$ as if they were independently distributed with marginal pdf

$$f(y \mid I_k = 1) = \frac{\mu(y;\xi)}{\int \mu(y;\xi) f(y)\, dy} f(y)$$

- The typical efficiency ordering:

Pseudo $<$ **Sample** $<$ **Full**

Sample likelihood estimation

- Sample likelihood has long history:
  - Patil and Rao (1978), Breslow and Cain (1988), Krieger and Pfeffermann (1992), Pf., Krieger and Rinott (1998), Pf. and Sverchkov (2009)
- But theoretical foundation has been less developed:
  - assuming n fixed as $N \to \infty$, PKR (1998) show pointwise convergence of joint pdf of responses to product of $f(y_k \mid I_k = 1)$
- Want theoretical results that account for dependence induced by design

Our contribution to sample likelihood estimation

- Bonnéry, Breidt, Coquet (2018, *Bernoulli*):
  - assume $\sqrt{n}$-consistent and asymptotically normal sequence of estimators of nuisance parameters ξ
  - often attainable via design-based regression: $\widehat{\xi}_{\mathrm{HT}}$
  - plug in $\widehat{\xi}_{\mathrm{HT}}$ to product of $f(y_k \mid I_k = 1; \theta)$:

$$\prod_{k \in s} \frac{\mu(y_k; \widehat{\xi}_{\mathrm{HT}})}{\int \mu(y; \widehat{\xi}_{\mathrm{HT}}) f(y; \theta)\, dy} f(y_k; \theta)$$

  - maximize with respect to θ to get $\widehat{\theta}_{\mathrm{SMLE}}$

Our contribution to sample likelihood estimation, II

- Consistency and asymptotic normality of $\widehat{\theta}_{\mathrm{SMLE}}$
  - assumptions are verifiable for some realistic designs
  - asymptotic approximations work well in simulations
- Asymptotic covariance matrix depends on
  - joint covariance matrix of score vector and $\widehat{\xi}_{\mathrm{HT}}$, estimated via design-based methods
  - information matrix for θ, estimated via model-based methods (plug SMLEs into analytic derivation)
- Design-based regression problem followed by classical likelihood problem

Approaches to testing

- **Approach 1:** Test for dependence on y_k of

$$\mathbb{E}\left[\Pi_k \mid Y_k = y_k\right] = \mu(y_k; \xi)$$

  – this is a regression specification test

  – parametric or nonparametric

- **Approach 2:** Test for a difference between design-weighted and unweighted ...

  – ... parameter estimates

  – ... probability density function estimates

  – ... cumulative distribution function estimates

Intuition of Approach 2

- Design-weighted corrects for ρ and targets f (perhaps inefficiently)
- Unweighted does not correct for ρ and targets ρf
- Difference between weighted and unweighted indicates $\rho \not\equiv 1$, so selection is informative

F-test based on difference in parameter estimates

- Consider the normal linear model with $\boldsymbol{x}_k$ and $\boldsymbol{x}_k$-by-design weight interactions (including intercept-by-weight):

$$\boldsymbol{Y}_s = \left[\boldsymbol{x}_k' \ \frac{1}{\pi_k}\boldsymbol{x}_k' \right] \begin{bmatrix} \theta \\ \boldsymbol{\gamma} \end{bmatrix} + \boldsymbol{\varepsilon}_s, \quad \boldsymbol{\varepsilon}_s \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 I\right)$$

  where $[\boldsymbol{x}_k']_{k\in s}$ is full-rank

- Algebraically, $\mathbb{E}\left[\widehat{\theta}\right] = \mathbb{E}\left[\widehat{\theta}_{\mathrm{HT}}\right] \Leftrightarrow \boldsymbol{\gamma} = \mathbf{0}$
- Test $H_0 : \boldsymbol{\gamma} = \mathbf{0}$ versus $H_a : \boldsymbol{\gamma} \neq \mathbf{0}$ via the usual F-test
  - DuMouchel and Duncan 1983; Fuller 1984

F-Test for gestational age example

- Full/alternative model: $Y_k \sim \mathcal{N}\left(\theta + \gamma(\pi_k^{-1}), \sigma^2\right)$
- Reduced/null model: $Y_k \sim \mathcal{N}\left(\theta, \sigma^2\right)$
- Test null hypothesis of non-informative selection:

```
> fit.full <- lm(GestAge ~ weight, data = birth)
> fit.reduced <- lm(GestAge ~ 1, data = birth)
> anova(fit.reduced, fit.full)
Analysis of Variance Table

Model 1: GestAge ~ 1
Model 2: GestAge ~ weight
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     89 1505.04
2     88  256.35  1    1248.7 428.66 < 2.2e-16 ***
---
```

Wald test based on difference in parameter estimates

- More generally, Pfeffermann (1993) derived the Wald-type test statistic,

$$W_N = \left(\widehat{\theta}_{\mathrm{HT}} - \widehat{\theta}\right)' \left\{-\widehat{J}^{-1} + \widehat{J}_{\mathrm{HT}}^{-1}\widehat{K}_{\mathrm{HT}}\widehat{J}_{\mathrm{HT}}^{-1}\right\}^{-1} \left(\widehat{\theta}_{\mathrm{HT}} - \widehat{\theta}\right)$$

where J and K matrices depend on

$$\pi_k^{-1},\ \mathrm{Var}\left(\frac{\partial \log f(y_k \mid \theta)}{\partial \theta}\right),\ \frac{\partial^2 \log f(y_k \mid \theta)}{\partial \theta\, \partial \theta'}$$

- Under the null hypothesis $\mathbb{E}\left[\widehat{\theta}_{\mathrm{HT}} - \widehat{\theta}\right] = \mathbf{0}$, W_N converges in distribution to a chi-squared distribution with degrees of freedom equal to dim(θ)

Test based on likelihood ratio

- Wald test requires considerable derivation
- Alternative test does not compare parameter estimates directly, but evaluates their likelihood ratio
  - unweighted log-likelihood ratio:

$$\mathrm{LR} = 2\left\{\ln \mathcal{L}(\widehat{\theta}) - \ln \mathcal{L}(\widehat{\theta}_{\mathrm{HT}})\right\}$$

  - weighted (pseudo) log-likelihood ratio:

$$\mathrm{LR}_{\mathrm{HT}} = 2\left\{\ln \mathcal{L}_{\mathrm{HT}}(\widehat{\theta}_{\mathrm{HT}}) - \ln \mathcal{L}_{\mathrm{HT}}(\widehat{\theta})\right\}$$

- (W. Herndon, 2014 CSU dissertation advised by Breidt and Opsomer, and joint with R. Cao and M. Francisco-Fernández)

- Under H_0 : non-informativeness, the LR test statistics converge,

$$\mathrm{LR} \xrightarrow{d} \sum_{i=1}^{p} \lambda_i Z_i^2, \quad \mathrm{LR}_{\mathrm{HT}} \xrightarrow{d} \sum_{i=1}^{p} \lambda_{\mathrm{HT},i} Z_i^2$$

where Z_i iid $\mathcal{N}(0,1)$ and λ_i, $\lambda_{\mathrm{HT},i}$ are eigenvalues of matrices involving

$$\pi_k^{-1}, \ \text{Var}\left(\frac{\partial \log f(y_k \mid \theta)}{\partial \theta}\right), \ \frac{\partial^2 \log f(y_k \mid \theta)}{\partial \theta \, \partial \theta'}$$

- Seems as bad as Wald, but . . .

Bootstrapping is easy

- Parametric bootstrap version of LR test statistic:
  - draw bootstrap sample from fitted density and construct LR test statistic B times
  - bootstrap p-value$= B^{-1}\sum_{b=1}^{B} \mathbb{1}\{\mathrm{LR}^{(b)} > \mathrm{LR}\}$
  - simple to implement: no information computations
- Both the linear combination of χ_1^2's and the bootstrap version work well in simulations
  - correct size under H_0
  - good power for a range of informative designs

Now consider nonparametric estimation and tests

- Nonparametric density estimation and testing
  - alternatives to "classic" design-weighted KDE
  - compare design-weighted KDE to unweighted KDE for testing?
- Nonparametric CDF estimation and testing
  - brief review of CDF estimation under informative selection
  - tests comparing design-weighted empirical CDF to unweighted CDF

Kernel density estimation under informative selection

- Bonnéry, Breidt, Coquet (2017, *Metron*)
- Under standard assumptions, unweighted KDE

$$\frac{1}{n}\sum_{k\in s}\frac{1}{h}K\left(\frac{y_k - y}{h}\right)$$

with kernel K, bandwidth h converges not to $f(y)$, but

$$\frac{\mu(y;\xi)}{\int \mu(y;\xi)f(y)\,dy}f(y) = \rho(y;\xi)f(y)$$

– usual $O(h^2)$ rate for bias, in estimation of ρf

– "usual" $O\left((Nh\int \mu f)^{-1}\right)$ variance

KDE under informative selection, continued

- Unweighted KDE converges to

$$\frac{\mu(y;\xi)}{\int \mu(y;\xi) f(y)\,dy} f(y) = \rho(y;\xi) f(y)$$

- **"Outer adjustment":** use unweighted KDE
  - estimate and remove ρ
  - or estimate and remove μ and $\int \mu f$
- **"Inner adjustment":** use weighted KDE
  - weights from inclusion probabilities regressed on y
  - or from design weights regressed on y

Outer = Inner for design-weighted KDE

- **"Outer adjustment":** Estimating μ and $\int \mu f$ via design-weighted nonparametric regression leads to

$$\frac{1}{\sum_{k \in s} \pi_k^{-1}} \sum_{k \in s} \frac{1}{h} K\left(\frac{y_k - y}{h}\right) \frac{1}{\pi_k}$$

- But this is just **"Inner adjustment"** using the original design weights
- This standard, design-weighted KDE is the baseline for comparison

Integrated MSE results with gestational age model

- $n = 90$, 1000 reps with 5-per-stratum in 18 strata

	$\mathbb{E}[\Pi \mid Y = y] = \mu(y;\xi)$	IMSE Ratio	$\mathbb{E}\left[\Pi^{-1} \mid Y = y\right] = \omega(y;\delta)$	IMSE Ratio
Outer	μ, ξ known	1.5	—	—
	ξ unknown	1.7	—	—
	misspecified μ	1.6	misspecified ω	1.6
	kernel reg.	1.0	kernel reg.	0.96
Inner	μ, ξ known	0.9	—	—
	ξ unknown	0.96	—	—
	misspecified μ	0.94	misspecified ω	0.93
	kernel reg.	1.4	kernel reg.	1.4

Testing for informativeness using KDE?

- KDE summary:
  - nonparametric outer adjustment works well
  - parametric inner adjustment works slightly better
- Design-weighted or adjusted KDE converges to f
- Unweighted KDE converges to ρf
- At a minimum, this is an exploratory tool that may suggest informativeness
- Formal testing is a subject of future work

CDF estimation under informative selection

- Bonnéry, Breidt, Coquet (2012, *Bernoulli*)
- Under mild conditions, the (unweighted) empirical CDF

$$\widehat{F}(\alpha) = \frac{\sum_{k \in U} \mathbb{1}_{(-\infty,\alpha]}(Y_k) I_k}{\mathbb{1}\left(\boldsymbol{I}_U = \boldsymbol{0}\right) + \sum_{k \in U} I_k}$$

converges uniformly in L_2:

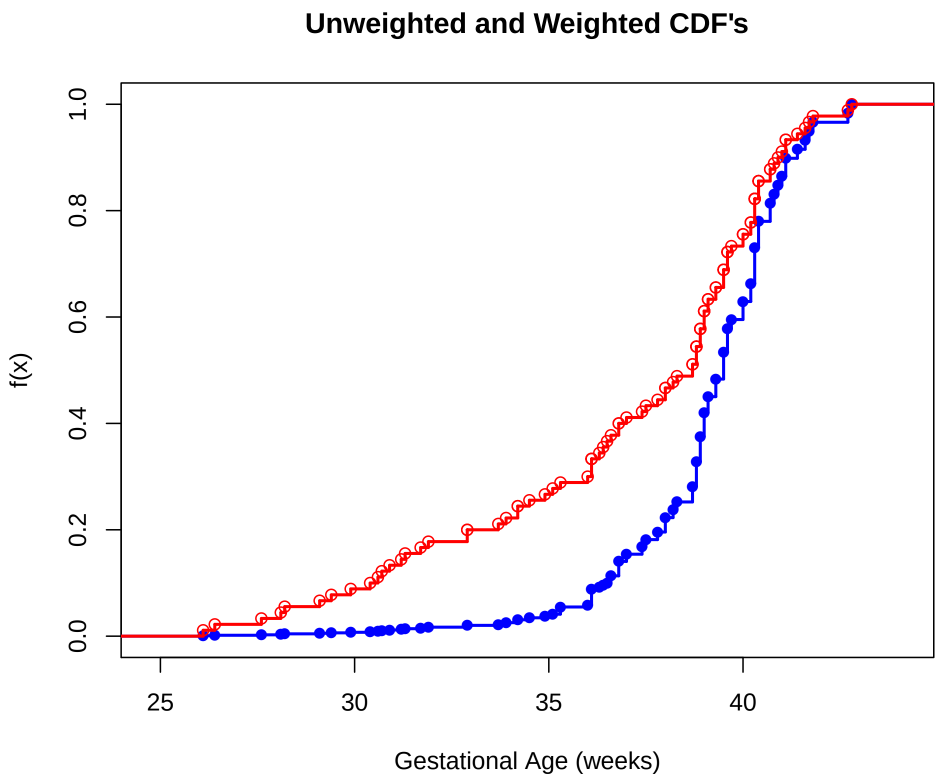
$$\sup_{\alpha \in \mathbb{R}} \left| \widehat{F}(\alpha) - F_\rho(\alpha) \right| = \|\widehat{F} - F_\rho\|_\infty \xrightarrow[N \to \infty]{L_2} 0$$

where the limit CDF is distorted by selection:

$$F_\rho(\alpha) = \frac{\int_{-\infty}^{\alpha} \mu(y;\xi) f(y)\, dy}{\int \mu(y;\xi) f(y)\, dy} = \int_{-\infty}^{\alpha} \rho(y;\xi) f(y)\, dy$$

Return to gestational age example

- Looks like informative selection: can we test?



Classical tests based on empirical CDFs

- Functional CLT for independent empirical CDFs:

$$D_n(\alpha) = \frac{\sqrt{n}}{2}\left\{F_n^{(1)}(\alpha) - F_n^{(2)}(\alpha)\right\}$$

converges in distribution to a **Brownian bridge:** zero-mean Gaussian process $\mathbb{G}_F$ with covariance function

$$\mathbb{E}\left[\mathbb{G}_F(s)\mathbb{G}_F(t)\right] = F(s \wedge t) - F(s)F(t)$$

- Kolmogorov–Smirnov two-sample test: $\|D_n(\alpha)\|_\infty$
- Cramér–von Mises two-sample test: $\int_{-\infty}^{\infty} D_n^2(\alpha)\, dF_n(\alpha)$, with $F_n = \psi F_n^{(1)} + (1-\psi)F_n^{(2)}$ for some $\psi \in [0,1]$

Adapting to the survey context

- Boistard, Lopuhaä, and Ruiz-Gazen (2017) develop functional CLT for

$$\sqrt{n}\left\{\frac{\sum_{k\in U}\mathbb{1}(Y_k\le\alpha)I_k\pi_k^{-1}}{\widehat{N}}-F(\alpha)\right\}$$

  via assumptions on

  – CLT for HT, to get finite dimensional distributions

  – higher-order inclusion probabilities, to get tightness

- Adapt and extend to weighted minus unweighted CDF:

$$T_N(\alpha)=\sqrt{n}\left\{\frac{\sum_{k\in U}\mathbb{1}(Y_k\le\alpha)I_k\pi_k^{-1}}{\widehat{N}_{\mathrm{HT}}}-\frac{\sum_{k\in U}\mathbb{1}(Y_k\le\alpha)I_k}{n}\right\}$$

  (Teng Liu, CSU PhD, 2019)

- **Result:** Under the null of no informative selection, $T_N(\alpha)$ converges in distribution to a **scaled Brownian bridge:** zero-mean Gaussian process $\mathbb{G}_F$ with covariance function

$$\mathbb{E}\left[\mathbb{G}_F(s)\mathbb{G}_F(t)\right] = C\left\{F(s \wedge t) - F(s)F(t)\right\}$$

where

$$C = \lim_{N\to\infty} \frac{n}{N^2} \sum_{k\in U} \mathbb{E}\left[\frac{1}{\Pi_k}\left(1 - \frac{N\Pi_k}{n}\right)^2\right]$$

Adapting to the survey context, III

- Estimate the scaling factor

$$C = \lim_{N \to \infty} \frac{n}{N^2} \sum_{k \in U} \mathbb{E}\left[\frac{1}{\Pi_k}\left(1 - \frac{N\Pi_k}{n}\right)^2\right]$$

using design-based methods:

$$\widehat{C}_{\mathrm{HT}} = \frac{n}{\widehat{N}_{\mathrm{HT}}^2} \sum_{k \in U} \frac{I_k}{\pi_k^2}\left(1 - \frac{\widehat{N}_{\mathrm{HT}}\pi_k}{n}\right)^2$$

Adapting to the survey context, IV

- Under probability-proportional-to-size sampling, the scale factor simplifies further: with $w_k = \pi_k^{-1}$,

$$\widehat{C}_{\mathrm{pps}} = (S_w/\bar{w})^2\,(n-1)/n \simeq (\mathrm{CV}_w)^2$$

- Kolmogorov–Smirnov test of informative selection:

$$\widehat{C}^{-1/2}\|T_n(\alpha)\|_\infty$$

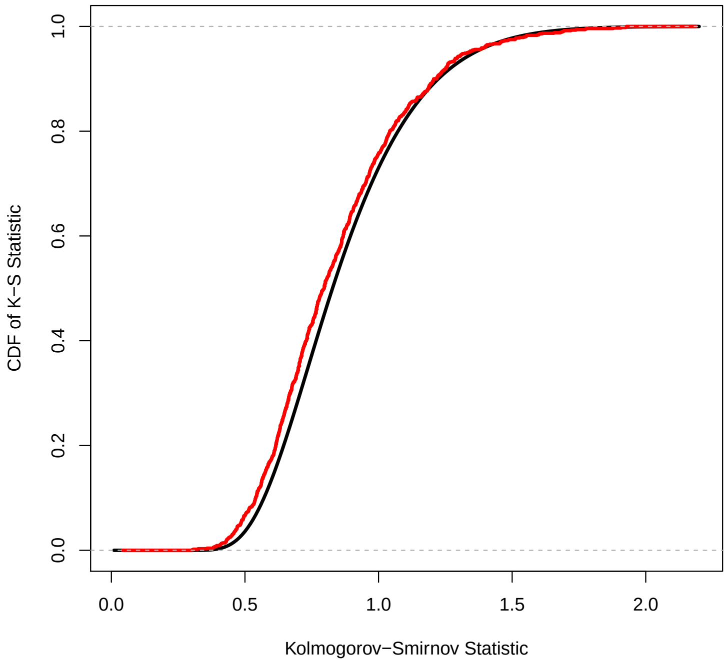
- Cramér–von Mises test of informative selection:

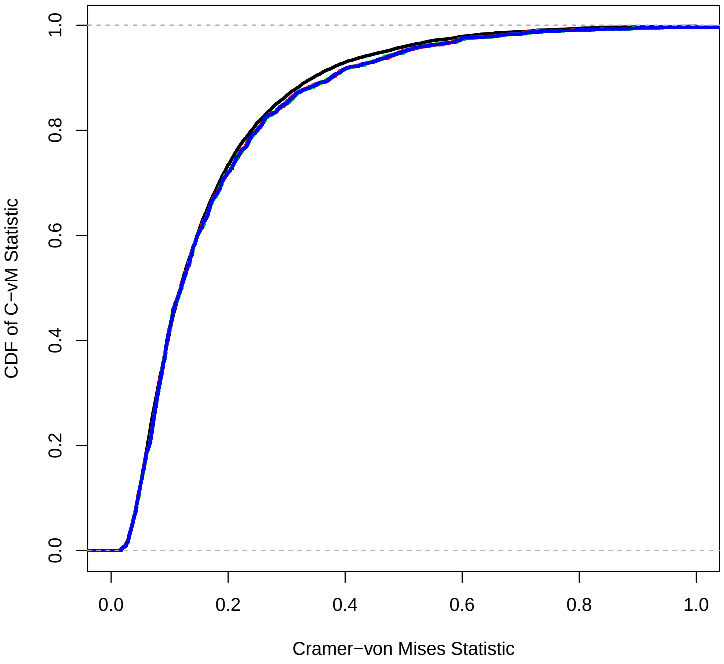
$$\widehat{C}^{-1}\int_{-\infty}^{\infty} T_n^2(\alpha)\,dH(\alpha),$$

with $H = \psi\widehat{F}_{\mathrm{HT}} + (1-\psi)\widehat{F}$ for some $\psi \in [0,1]$

Test statistic distributions for gestational age

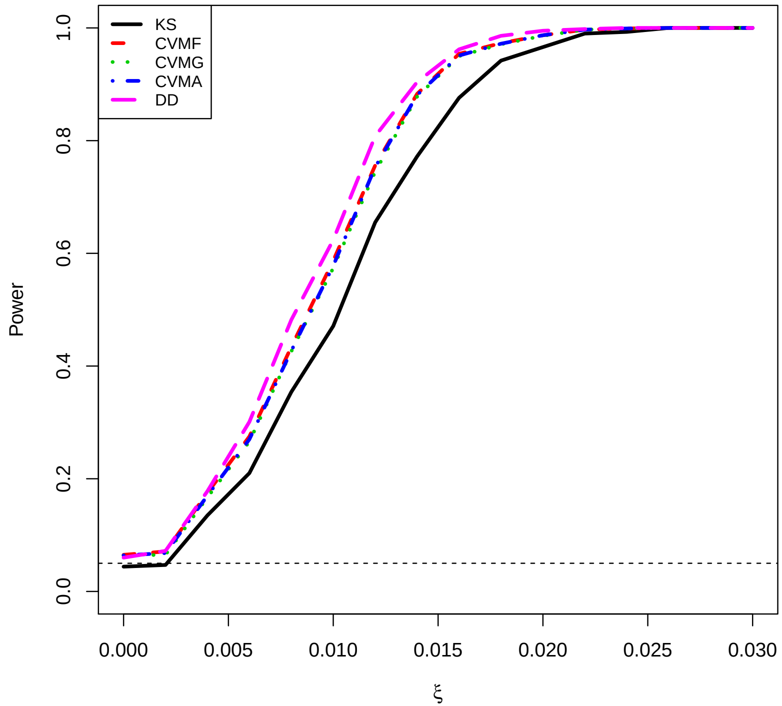
- Asymptotic distribution and empirical distribution of K–S and C–vM, with $n = 300$ and 1000 reps





Power for gestational age simulation

- Empirical $\xi = 0.175$ in $Y_k \mid (I_k = 1) \sim \mathcal{N}\left(\theta - \xi\sigma^2, \sigma^2\right)$
- Choose grid of $\xi \in [0, 0.03]$; use $n = 300$ and 1000 reps each



A different example!

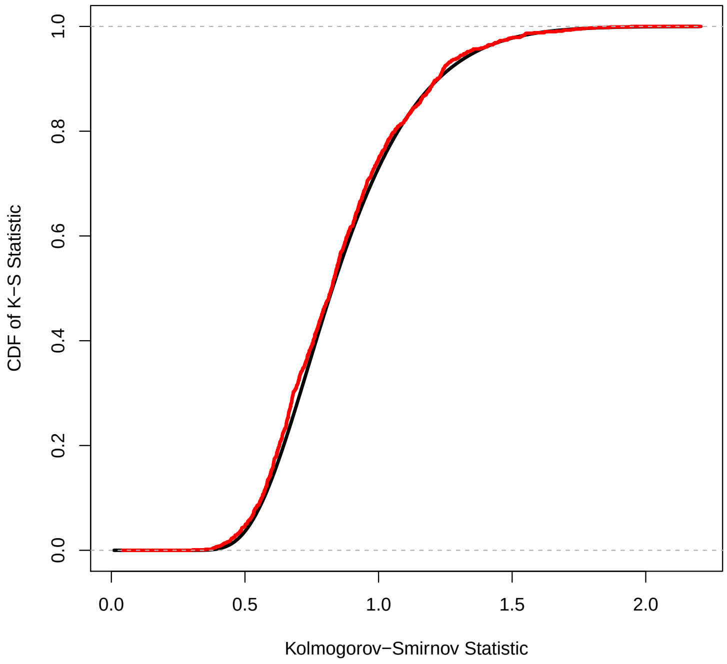
- Suppose Y_k are iid location-scale t_ν:

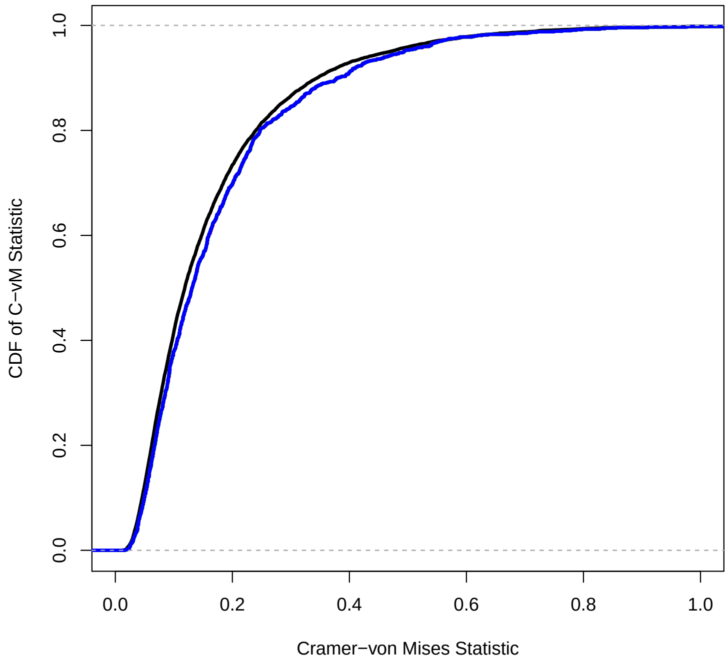
$$Y_k = \theta + \sigma \frac{Z_k}{\sqrt{V_k/\nu}} \sqrt{\frac{\nu - 2}{\nu}} = \theta + \sigma_k Z_k,$$

  $\{Z_k\}$ iid $\mathcal{N}(0,1)$ independent of $\{V_k\}$ iid χ^2_ν
- Informative Poisson sampling with $\pi_k \propto \sigma_k$
  - minimizes design-model variance of HT estimator
- $\sigma_k \to \sigma$ as $\nu \to \infty$, and informativeness disappears

Test statistic distributions for location-scale t_ν

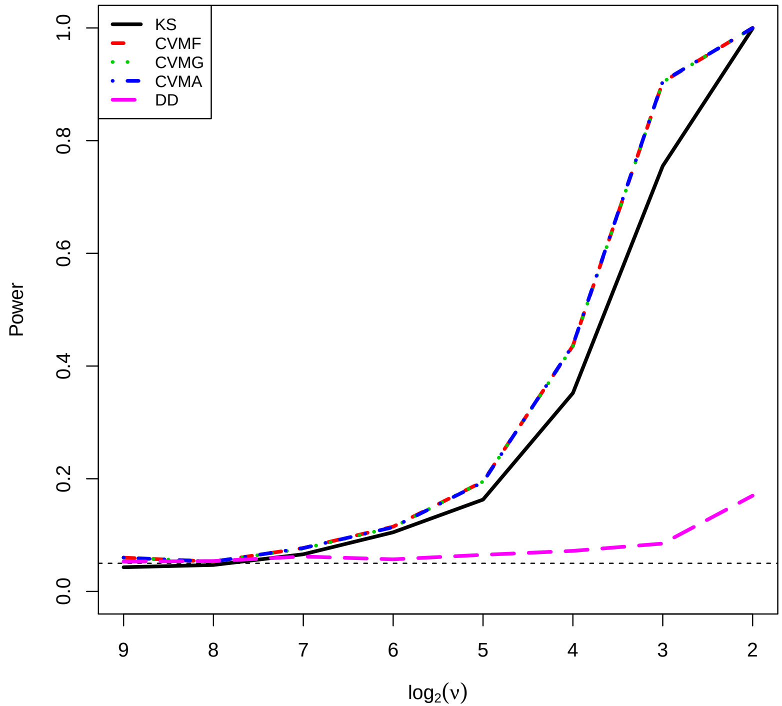
- Asymptotic distribution and empirical distribution of K–S and C–vM, with $n = 300$ and 1000 reps





Power for location-scale t_ν simulation

- Choose $\nu = 2^2, 2^3, \ldots, 2^9$; use $n = 300$ and 1000 reps each
- DD test gets some "lucky" power at low df due to random variation



Lucky power?

- Weighted and unweighted estimators have the same mean
- At very low degrees of freedom, HT is (particularly) highly variable
- Difference between weighted and unweighted is large due to chance variation
- DD correctly rejects by incorrectly assuming large difference is a difference in the mean

Summary

- Informative selection is pervasive
- Strategy of comparing weighted to unweighted works broadly:
  - parametric, from linear models to likelihood ratios
  - nonparametric, from kernel density estimation to classic two-sample tests
- Design-weighted estimation is a "safe" and readily-available solution
- Sample likelihood approach is a viable alternative

THANK YOU

- Thank you for your attention
- Thanks to Matthieu, Guillaume, and Yves for a wonderful conference!